DOCUMENT RESUME

ED 277 277 FL 016 341

AUTHOR van Weeren, J.; Theunissen, T. J. J. M.
TITLE Testing Pronunciation: An Application of Generalizability Theory.
INSTITUTION Centraal Inst. voor Toetsonwikkeling, Arnhem (Netherlands).
PUB DATE 86
NOTE 18p.; Table III in the appendix contains small print.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Audiotape Recordings; *Error Patterns; French; *Generalizability Theory; German; Holistic Approach; *Language Tests; *Pronunciation; Second Language Instruction; Second Language Learning; Test Construction; Test Format; Test Items; *Test Reliability; Test Validity

ABSTRACT

Pronunciation is regarded as a valuable subskill in foreign language teaching and testing. Its quality is commonly assessed in a global way by having examinees read aloud. An atomistic test is a more systematic and explicit approach. Such a test would consist of about 40 items, use recorded performances, and draw on an inventory of pronunciation problems to locate error patterns in students' speech. A study of such a test using Dutch learners of French and German as a second language found it to be reliable for these groups. Further research into the validity of this kind of test is needed, since there are factors such as prosody, liveliness, and overall articulation that would figure in a global rating but are not addressed by this rating system. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED277277

TESTING PRONUNCIATION:
AN APPLICATION OF GENERALIZABILITY THEORY*

J. van Weeren and T.J.J.M. Theunissen
National Institute for Educational Measurement (Cito),
The Netherlands.

Pronunciation is regarded as a valuable subskill in foreign language teaching and testing. Its quality is commonly assessed in a global way by having testees read aloud. A more systematical and explicit approach to evaluation is proposed. The reliability of this so called 'atomistic' approach is compared with the reliability of a traditional 'holistic' rating procedure.
Generalizability theory is applied in order to comprise all relevant factors in one single psychometric model. Decision studies are carried out to settle the effects of number of raters and test length.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
J. VAN Weeren
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
■ This document has been reproduced as received from the person or organization originating it.
□ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

There are good reasons for not neglecting pronunciation as an important subskill in foreign language teaching and testing. First of all, with a good pronunciation one can make oneself better understood. It gives one's oral production a certain redundancy and this can help a speaker to get his message across more easily. Secondly, pronunciation is quite attractive as a learning objective because of its high pay-off. The number of sounds, sound clusters and intonation patterns in a language is finite, just like the alphabet. Once the system has been mastered it can be applied in one's future performance, thus giving it fundamentally an infinite scope. Thirdly, a deviant pronunciation means that one is immediately 'marked' abroad. A 'foreign accent' is often used as the means to distinguish non-natives from native speakers. There can be some discussion about the proposal that foreigners should cherish their rôle as non-natives. However, one can think of learners who would prefer a native-like pronunciation, for example those with a

* We would like to thank Jan-Arjen Mondria, who as a guest-worker has carried out the reported experiment for French.

FL016341

particularly strong integrative motivation (Gardner and Lambert 1972) and one can imagine situations in which one wishes to avoid immediate identification as a foreigner.

If pronunciation is worth teaching, it is worth evaluating. For this evaluation one needs a speech sample of the testee. This can be obtained by involving him in a more or less spontaneous conversation or by asking him questions about a text he has read before, about his hobbies, his plans for the future etcetera. However, it is difficult to separate pronunciation from other aspects of the testee's oral production, such as fluency, vocabulary, grammatical accuracy and judge it independently (Clark and Swinton 1980). If the testing focus is on pronunciation, it is a commom practice to have the testee read aloud a few lines or paragraphs of a written text. The advantage of this procedure is that one can obtain comparable speech samples from different testees. It cannot be denied, however, that this elicitation procedure evokes an unknown variable: the ability to read aloud. The impact of this variable is unclear. As far as we know, no relevant research has been carried out in order to determine its effect. It could cause severe problems with testees who are to some extent dyslectic. Nevertheless, reading aloud is very common in language teaching and usually testees are familiar with it. Moreover, we cannot neglect the advantage of obtaining comparable speech samples. Having them collected is necessary in order to determine the quality of testing procedures.

Normally, rating is done by giving the testee a mark for the overall impression of his pronunciation. Rating can be based on a four-, five- or six-point-scale. Every point on the scale is defined by a description of one or two prominent characteristics of the testee's performance, like 'occasional phonemic errors, but generally comprehensible' or 'many phonemic errors: very difficult to perceive meaning' (Clark 1972: 93). The selection of a certain mark or scale point can be both supported and motivated by a written record made during the testee's performance in which pronunciation errors are successively noted down. The accountability of this rating procedure can be improved by working out such records in a systematical way and by letting the final mark depend on it entirely. Furthermore, a systematic and complete record can be used for diagnostic purposes. Weak points in a testee's pronunciation can be subjected to subsequent training. Proposals for setting up such records are found in several testing guidelines (e.g. Harris 1969: 86-87, Heaton 1979: 84-86, and Madsen 1983: 66-68).

These procedures have in common that in a running text or in a list of unconnected sentences possible pronunciation problems, including stress and intonation problems, are marked; for example (Harris 1969: 86):

While Mr. Brown read his newspaper,	primary stress
his wife finished packing his clothes	voiced final consonant(s)
for the trip. The suitcase was al-	
ready quite full, and she was having	vowel quality
a great deal of difficulty finding	primary stress
...	

An unmarked version of the test is read aloud by the testee and the tester ticks off the places where an error is made. The advantage of this procedure is obvious: a judgement of a testee's pronunciation can be accounted for by means of explicitly defined categories. Therefore the procedure is more transparant. It should be preferred to the traditional method from a test theoretical point of view.

Statement of the problem

We decided to examine the reliability of a rating procedure for the evaluation of pronunciation by which only marked elements in a text read aloud are judged. The realization of each element could be judged either as correct or as incorrect. The reliability of this procedure was to be compared with the reliability of the common practice when an overall impression mark is given for the quality of the testee's pronunciation after reading a text aloud. We will call the first method 'atomistic' and the second 'holistic'.

Pretest

First of all an inventory was made of common pronunciation errors in German and French as spoken by Dutch people. Such inventories can be extracted from several textbooks for the teaching of the pronunciation of German and French. A rough outline of our inventories is given in the left-hand columns of tables

I and II in the appendix.

Two texts were selected and adapted slightly. On average every 5th or 6th word (German) and every 8th or 9th word (French) contained a pronunciation problem. The German text was read aloud by 10, the French text by 14 adults at normal speed. All testees had learned the respective foreign language in secondary school. The testees' taperecordings were marked by a panel of fully qualified teachers, 7 for both German and French, after a short instruction. The taperecordings were presented only once, without stops or replays. The raters were to mark the separate items on a rating sheet of the type that was outlined above. Only incorrect realizations of elements in the texts were marked, so that for each testee and rater the number of pronunciation errors could be calculated.

However, unmarked items are ambiguous with regard to their interpretation. An unmarked item can mean two things:

1 no pronunciation error has been made

2 a pronunciation error has not been registered.

Items that are hardly marked by the raters at all are of no interest. The testees' pronunciation is either correct or errors are not evident enough. Such items are superfluous if the testees are to be discriminated on the basis of the quality of their pronunciation. Items that are marked as incorrect to a substantial extent are more important in this respect.

For a good item, i.e. an item with a certain discriminative power, registration of errors is a necessary, but not a sufficient condition. It is also necessary that a vast majority of raters should react in the same way on the same item and the same testee. Ideally, every rater should mark an item as incorrect each time that it is realized incorrectly and do nothing if it is realized in a correct way. In that case the procedure is absolutely reliable. However, as a rule absolute reliability cannot be attained. Factors such as lack of power of perception or flagging concentration among the raters will prevent this. Nevertheless a certain minimal requirement of reliability must be met, if the rating procedure is to be worthwhile.

In order to select successful items for the testing procedure under investigation we set the following conditions:

1 a majority of raters (5 out of 7) should register an error made by a certain testee on a specific item;

2 the raters should show a certain measure of agreement in their ratings of all the testees on a particular item. A minimum value of .70 for an

---

4

interrater reliability coefficient was set.

As an index for agreement we used percentage of agreement uncorrected for chance (PUC). The agreement among n raters for any subject is expressed in the proportion of agreeing pairs of raters for all n (n-1) pairs of raters, and may be interpreted as the chance that a randomly selected rater agrees with a previously randomly selected rater.

For German, 13 out of 41 items met both criteria. For French this was the case for 15 out of 40. All the German items concerning vowels and diphthongs were to be discarded, as were those concerning the incorrect voicing of [s] caused by assimilation, the aspiration of unvoiced plosives, the realization of [R] as [X] and the pronunciation of [w] as [u̯] or [u]. The German test was revised in such a way that all the fitting pronunciation problems appeared in the text twice. Two items concerning the glottal stop were added. The new test consisted of 24 items. It is contained in the appendix (III). In the French test most of the vowel items on diphthongizing, (lack of) nasalization and reduction turned out to be unsatisfactory too, as did the items aiming at the devoicing of voiced consonants. Still, most of the pronunciation problems were covered again by the new test. On the basis of the recorded material other difficult sounds in the text were selected in order to construct alternative items. The revised version of the text again consisted of 40 items. It is also contained in the appendix (IV). An item on sentence intonation was added.

Consequently the main difference between the German and the French test was that the first was revised in a strictly mechanical way: only pronunciation problems represented by items that passed the pretest were selected, whereas the French test was adjusted with the help of the actual taperecordings. Almost all the original pronunciation problems were preserved in the definitive version.

Test and analysis of data

For the trial of the definitive tests new taperecordings were made. The text was read aloud by pupils in secondary schools who had learned German for 2-4 years and French for 3-5 years. In order to attain a normal rate of speech, the supervisor first read out a few (other) lines himself by way of an example. There were 26 testees and 24 items for German and 19 testees and 40 items for French. The taperecordings were rated by 14 fully qualified teachers of German and 15 fully qualified teachers of French, again after a short instruction. The rating procedure was carried out in the same way as in the pretest.

The purpose of the data analysis was to estimate the reliability of the atomistic procedure. The procedure is, as it were, a conjunction of the effects of testees, items, raters and their interactions. Classical indices of reliability are of limited use, since to a large extent they ignore the multiple sources of variation in the observations. Generalizability (G) theory takes this multiplicity explicitely into account. In a G-study, one decides in advance what factors have a known or hypothesized influence on the observations, and which of these factors are to be included in the study, as in a factorial analysis of variance design. Furthermore, the plausible assumption is made that factual conditions resulting from those decisions are representative random samples from their respective universes of possible conditions, or, alternatively, some factors may be regarded as fixed, i.e., not (randomly) sampled from a universe. In the present study, for obvious reasons relevant factors were considered to be testees, items and raters.

In a testees x items x raters design, variation in the observations stems from these factors and their interactions. They are sources of variance and their effects are expressed in variance components (for more details, see Bolus et al. 1982, Brennan 1983). In this design, we hope to find relatively low values for the interaction components, since they contribute to the error term a relatively high value for the testee component, combined with a low value for the rater component. The value of the item component is not important, as long as all testees are set the same items (as is usually the case in a test). The basic idea of a reliability index is to form the ratio of true variation and observed variation (true variation to which all sources of error-variation are added).

Since in this study our object of measurement is the testee, the index for generalizability is

$$\frac{\sigma^2_t}{\sigma^2_t + \frac{\sigma^2_{tr}}{n_r} + \frac{\sigma^2_{ti}}{n_i} + \frac{\sigma^2_{tir}}{n_r \times n_i}}$$

where σ^2_t is the variance component for testees, σ^2_{tr}, σ^2_{ti} and σ^2_{tir} are interaction components for testees with raters, testees with items and testees with items and raters, respectively, and n_r = number of raters, n_i = number of items.

It is clear from the above formula, that when all interactions are zero, the index has value 1, its maximum. That this should be so, is obvious when one remembers the meaning of these components.

It is also clear, that interaction components have to be interpreted with the appropriate sizes of samples in mind. Lengthening the test or increasing the number of raters produces higher reliability, a standard result in classical test theory.

Having assessed the various components, one may systematically study their effects in a so-called D(decision)-study by varying the numbers n_r and n_i. One may reduce them to simulate practical class-room situations, or increase them to assess the effect of more raters or items (more money or more testing-time!) on the reliability of the measurements.

Results for French

We will report the French results first. Table 1 gives the variance components.

Table 1, variance components, French test

testees	0.011
raters	0.005
items	0.020
testees x raters interaction	0.001
testees x items interaction	0.075
raters x items interaction	0.014
residual	0.066

In situations like this (large number of items) the relatively high components for items and testees x items can be expected and are quite normal. The relative importance of the testees x items components in the denominator of the given formula is small, since n_i is fairly large, and therefore also $n_r \times n_i$. The rater component is relatively low, which indicates the use of the same standard: all raters are equally severe or permissive. The testees x raters component is small, which means that agreement exists over all subjects. The relatively larger raters x items interactions indicate a certain lack of consistency over the items. Because of small σ^2_{tr} and large n_i, generalizability indices are fairly high. The test discriminates moderately well:

Table 2, generalizability coefficients, French test: 1 rater/2 raters

20 items random	.58/.66
30 items random	.66/.73
40 items random	.72/.78

These coefficients show the effects of testlength and number of raters. Adding 10 items to a test consisting of 20 or 30 items has roughly the same effect as the addition of a second rater. If testing-time is restricted because of a large number of testees, then bringing in a second rater is recommended if reliability is to be increased. However, if the man-hours of raters are limited or if the addition of a second rater is too difficult to organize, one should lengthen the test.

The coefficients in table 2 are based on the assumption that the tests of

different lengths consist of a random selection of 'possible' pronunciation items. However, the 40-item-test as used by us can also be considered as a unique and definitive pronunciation test. In that case the items are considered as 'fixed' and generalizability coefficients are then based on the actual quality of the developed test as a whole (table 3):

Table 3, generalizability coefficients, French test: 1 rater/2 raters

40 items fixed	.86/.92

As can be seen from table 3, in this 'fixed' test the addition of a second rater adds only about .06 to the value of the coefficient.

Results for German

The results of the German test are presented below, first of all the variance components.

Table 4, variance components, German test

testees	0.003
raters	0.002
items	0.028
testees x raters interaction	0.001
testees x items interaction	0.029
raters x items interaction	0.034
residual	0.110

Compared with the French results the between-testees variance was relatively low. The German test could not identify as many differences among the testees as the French test could. The group of testees was very homogeneous as can be concluded from the results of the global rating (see below). Again, items differed greatly as to difficulty. The relatively high raters x items interaction indicates that the raters did not mark individual items in the same way. Furthermore, the residual variance is high compared with the French results. Lack of between-testees variance is reflected by the generalizability

coefficients (Table 5):

Table 5, Generalizability coefficients, German test: 1 rater/2 raters

20 items random	.27/.38
30 items random	.35/.47
40 items random	.41/.53

Compared with the French results, the addition of a second rater is more effective as the generalizability coefficients are lower.
The German test consisted of 24 items and for the comparison with French the generalizability coefficients were computed for 40 items. One of the advantages of the generalizability theory becomes apparent here: the coefficient can be computed in order to compare tests of different lengths. If items are considered 'fixed', which means that the test is lenghtened with items of the same average quality we get the following results:

Table 6, generalizability coefficients, German test: 1 rater/2 raters

40 items fixed	.51/.67

Again, adding a second rater has much more effect this time, whereas lengthening the test, for example up to 50 items, is not very effective; only a small increase of the coefficients (.02-.04) will be the result.

Setting a criterion

So far we have only compared the results of the French and German tests. It would be helpful to know if the new atomistic procedure would entail an improvement of traditional holistic practice. In order to answer this question the taperecordings were played once more and the raters were asked to give the testees a traditional impression mark on a twenty-point-scale (that is: the ten-point-scale with half points as is commonly used in the Netherlands).
Table 7 presents the variance components and the generalizability coefficients.

Table 7, traditional holistic rating, variance components and G-coefficients

	French	German
testees	1.9117	.8758
raters	.2592	.2287
residual	.5098	.5287
1 rater/2 raters	.79/.88	.62/.76

Again, the homogeneity of the German group of testees compared with that of the French group becomes apparent. Between-raters variance plays a certain part, but is more or less overruled by the between-testees variance. A comparison of table 7 with tables 3 and 6 shows that the French test is superior to the global rating and that the German test is not.

Conclusions and discussion

An atomistic test for pronunciation can further the transparency and accountability of the evaluational procedure. It should consist of about 40 items and be designed on the basis of recorded performances. With an inventory of pronunciation problems as a starting point one should try to find regular pitfalls in a text as it is spoken. However, if the same criteria for item quality are applied on the definitive French test as with the pretest items concerning vowels and diphthongs are less fitting (see appendix IV). Inadmissable diphthongizing can be ruled out as well as items on devoicing of voiced plosives.

In this study we focussed on the reliability of a rating procedure that tried to make the evaluation of pronunciation more explicit and transparent. The pronunciation problems in the items can be considered symptomatic for the quality of the French and German as spoken by Dutch learners.

It should be stressed, however, that there could be factors involved in a global rating that are not covered by the test, such as prosody, liveliness and overall articulation. Further research into the validity of the procedure should be carried out. Last but not least: the validity of evaluating pronunciation on the basis of a text read aloud has not been determined yet. This question would require a very sophisticated and inventive design.

REFERENCES

Bolus, R.E., Hinofotis, F.B. and K.M. Bailey 1982.
An Introduction to Generalizability Theory in Second Language Research.
Language Learning 32: 245-58.

Brennan, R.E. 1983.
Elements of Generalizability Theory. American College Testing Program.
Iowa City.

Clark, J.L.D. 1972.
Foreign Language Testing: Theory and Practice.
CCD, Philadelphia.

Gardner, R.C., and W.E. Lambert 1972.
Attitudes and Motivation in Second Language Learning.
Rowley (Mass.).

Harris, D.P. 1969.
Testing English as a Second Language. McGraw-Hill, New York.

Heaton, J.B. 1979.
Writing English Language Tests. Longman, London.

Madsen, H.S. 1983.
Techniques in Testing. Oxford University Press.

Newton, J.B. 1979.
Writing English Language Tests. Longman, London.

APPENDIX

Table I, German

pronunciation errors	number of items in pretest	number of items in test
1 vowels/diphthongs - dipthongizing - nasalization - quality	12	-
2 fricatives - voicing/unvoicing - pronunciaton of ʃ, ç, X	16	12
3 plosives - voicing/unvoicing - aspiration	4	2
4 rest - pronunciation of l, R, w - realization of -en # - 'Sprossvokal' - stress	9	10 -glottal stop
Total	41	24

Table II, French

pronunciation errors	number of items in pretest	number of fit items after pretest	number of items in test	number of fit items after test
vowels/diphthongs - no nasalization - diphtongizing - reduction	9	1	8	3
2 consonants - unvoicing - pronunciaton of ʃ, 3, l, R, ɥ - pronunciation of /h/	17	5	13	9
3 rest - stress - liaison/enchaînement - addition/elimination of sounds - pronunciation of specific words	14	9	19 -intonation	15
Total	40	15	40	27

III GERMAN TEST

Achim war von seinem Vetter
eingeladen worden, das
Wochenende in dessen Landhaus zu
verbringen. Als er am
Sonnabend ankam, waren schon viele
Gäste da. Abends, nach dem Essen,
saß man vergnügt zusammen, man
trank sehr viel, und nicht nur Tee
und Milch. Gegen Ende des Festes
kam Achims Vetter und sagte: "Mein
lieber Achim, du mußt im
Gespensterzimmer schlafen.
Alle andern Zimmern sind belegt."
"O, Achim", riefen die Damen,
"haben Sie denn gar keine Angst?
Sie wissen doch, in diesem
Zimmer geht die alte Tante um,
die vor acht Jahren dort
Selbstmord verübt hat." "Ach,
Unsinn, woher weiß man denn das?
Und Was ist dabei?" Kurze
Zeit später liegt Achim im Bett.
Er is aber doch ein bißchen
unruhig; einen revolver hat er
auf den Nachttisch gelegt. Im
Halbschlaf sieht er plötzlich
fünf kleine schwarze Finger, die
sich langsam am Fußende des
Bettes bewegen ... Er macht die
Augen weit auf. Er täuscht sich
nicht ... jetzt sind es sogar
zehn geworden! Achim stütz sich
auf. "Lassen Sie das!" sagt er kalt."
Zeigen Sie Ihr Gesicht, oder ich
schieße!"

1 ☐ no glottal stop
2 ☐ no primary stress
3 ☐ g wrongly voiceless
4 ☐ s wrongly voiceless
5 ☐ v wrongly voiced
6 ☐ 'Sproßvokal'
7 ☐ v wrongly voiced
8 ☐ no 'Achlaut'
9 ☐ no 'breiter Zischlaut'
10 ☐ no tense l
11 ☐ -n not spoken
12 ☐ g wrongly voiceless
13 ☐ no 'Achlaut'
14 ☐ 'Sproßvokal'
15 ☐ no glottal stop
16 ☐ z not as ts
17 ☐ no "Ichlaut'
18 ☐ no primary stress
19 ☐ no 'breiter Zischlaut'
20 ☐ s wrongly voiceless
21 ☐ no 'Ichlaut'
22 ☐ -n not spoken
23 ☐ no tense l
24 ☐ z not as ts

IV FRENCH TEST

Cinq minutes avec Jan Janssen.

Ancien champion du monde et vainqueur du Tour de France en mil neuf cent soixante-huit, Jan Janssen était hier dans la caravane du Tour de France, accompagné de son ami, Henri Anglade. A trente-six ans, Jan Janssen n'a pas beaucoup changé et il dit toujours venir en France avec beaucoup de plaisir.

'Je suis venu voir le Tour de France pour me changer un peu les idées. Cela fait toujours plaisir de revoir des anciens amis, et une ambiance que j'ai si longtemps connue. J'aime ce spectacle.'

'Qu'est-ce que vous faites encore pour le cyclisme?'

'A vrai dire, plus grand-chose. J'ai cessé de travailler pour la radio et les journaux en Hollande et j'ai mis fin aussi à mes activités d'organisateur de courses. Il ne me reste plus que ma marque de vélos, qui sont fabriqués en Belgique, par un bon petit constructeur et qui se vendent bien. Je ne fais que des vélos de course de qualité. Dix mille exemplaires ont été produits l'an dernier. Je suis content.'

'Le cyclisme actif ne vous manque pas?'

'Non, pas tellement. Mes affaires m'occupent beaucoup, je voyage beaucoup et je n'ai presque plus le temps d'assister aux courses. Aujourd'hui, je suis un peu en vacances. Cependant, je continue de prendre ma bicyclette mauve deux ou trois fois par semaine, tôt le matin, pour conserver la ligne.'

'Comment se porte le cyclisme hollandais?'

'Bien, très bien même, et nous possédons avec Kuiper un beau champion du monde. Voilà un garçon sérieux et courageux qui mérite absolument les résultats qu'il obtient. Zoetemelk reste égal à lui-même. Quant à Schuiten je pense qu'il manque de volonté. Il aurait dû faire de son mieux, dans la montagne, pour apprendre le métier.'

'Et le cyclisme en général?'

'Je constate que l'ambiance n'est plus ce qu'elle était

1	☐ [ʃ] instead of [sj]
2	☐ not clearly pronounced as [s]
3	☐ not pronounced as [jɛR]
4	☐ no liaison
5	☐ [ɪ̯] incorrect
6	☐ vowel reduction to [ə]
7	☐ [sj] instead of [ʃ]
8	☐ vowel diphthongized
9	☐ insufficient nasalization
10	☐ pronounced as [-kəl]
11	☐ not voiced
12	☐ not voiced
13	☐ no primary stress
14	☐ false liaison
15	☐ pronounced as [kuR] or [kuRsəs]
16	☐ vowel diphthongized
17	☐ pronounced as [-dɑ̃] or [-dənt]
18	☐ pronounced as [kwa-]
19	☐ no liaison
20	☐ not pronounced as [-ɑ̃]
21	☐ no voiced liaison
22	☐ not voiced
23	☐ no nasalization
24	☐ vowel reduction to [ə]
25	☐ not voiced
26	☐ no nasalization
27	☐ 'h' pronounced
28	☐ not voiced
29	☐ not voiced
30	☐ vowel diphthongized
31	☐ 'lui' pronounced as 'Louis'
32	☐ no nasalization
33	☐ [R] improper (as semivowel or fricative)
34	☐ no tense l
35	☐ no 'enchaînement'

il y a dix ans. Les coureurs sont moins gais, plus nerveux. Ils veulent réussir trop vite sans prendre le temps d'apprendre le métier à fond. Je pense qu'ils feront des carrières moins longues que celles des coureurs de ma génération, ce qui est regrettable.'

36 ☐ not voiced
37 ☐ no tense l
38 ☐ not voiced
39 ☐ not voiced
40 ☐ pronounced as [bəl]